Spatio-temporal spike and slab priors for MMV problems

Michael Riis Andersen, Ole Winther & Lars Kai Hansen
DTU Compute, Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark
Email: {miri, olwi, lkh}@dtu.dk

***Abstract*—We are interested in solving the multiple measurement vector (MMV) problem for instances, where the underlying sparsity pattern exhibit spatio-temporal structure motivated by the electroencephalogram (EEG) source localization problem. We propose a probabilistic model that takes this structure into account by generalizing the structured spike and slab prior and the associated Expectation Propagation inference scheme. Based on numerical experiments, we demonstrate the viability of the model and the approximate inference scheme.**

I. Introduction

The multiple measurement vector problem (MMV) [1] is given by:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times D}$ is the forward matrix, $\mathbf{Y} \in \mathbb{R}^{N\times T}$ is the measurement matrix, $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_T\end{bmatrix} \in \mathbb{R}^{D\times T}$ is the desired solution and $\mathbf{E} \in \mathbb{R}^{N\times T}$ is a matrix of corruptive noise. We are interested in finding sparse solutions to eq. (1) in the ill-posed regime, where $N < D$. Furthermore, the sparsity pattern of $\mathbf{X}$ is assumed to have certain structural properties. In particular, we are considering problems where the sparsity pattern exhibit spatio-temporal structure as in EEG source localization [2], [3] or in background subtraction in computer vision [4]. Let $\mathbf{z}_t$ be an indicator for the support of $\mathbf{x}_t$, i.e. $\mathbf{z}_t = \mathbb{I}[\mathbf{x}_t \neq 0]$, then $\mathbf{z}_t$ is assumed to be spatially correlated. Furthermore, we assume that the support vectors $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T$ slowly evolve through time as well - rendering the joint sparsity assumption invalid [5].

The main contribution of this work is to propose a model for spatio-temporal sparsity patterns by extending the structured spike and slab prior [6] to account for temporal evolution of the sparsity pattern as well. Furthermore, we demonstrate the benefits of the model through numerical experiments.

A. Related work

The field of structured sparsity has received a great deal of attention in the recent years. In this section we highlight some of the related work, but this list is by no means exhaustive. The LASSO-community have introduced the Group and Graph LASSO methods, which generalize the standard ℓ_1-norm minimization approach to promote different kinds of structured sparsity [7]. In the probabilistic setting, the standard workhorse for sparsity is the so-called spike and slab prior [8]. This has also been generalized to model group sparsity [9] and cluster sparsity [10]. In the context of compressed sensing [11], Cevher et al. [4] used a Markov random field to enforce spatially correlated sparsity patterns, whereas Ziniel et al. used binary Markov chains to model temporally correlated sparsity patterns [12].

II. The structured spike and slab prior

In this section we briefly introduce the conventional spike and slab prior [8] and the structured spike and slab prior [6] before we move on to the spatio-temporal spike and slab prior on the next section. The conventional spike and slab prior decomposes each $x_{i,t}$ as a product of a binary variable $z_{i,t}$ and a real number $c_{i,t}$, i.e. $x_{i,t} = z_{i,t}c_{i,t}$, where $z_{i,t} \sim \text{Ber}(p_0)$ and $c_{i,t} \sim \mathcal{N}(0, \tau_0)$ for $i \in \{1, 2, .., D\}$ and $t \in \{1, 2, .., T\}$. The structured spike and slab prior generalized this formulation by imposing structure on the binary variable for each time t as follows

$$p(\mathbf{z}_t|\phi(\boldsymbol{\gamma}_t)) = \prod_{i=1}^{D} \text{Ber}\left(z_{i,t}|\phi(\gamma_{i,t})\right), \tag{2}$$

$$p(\boldsymbol{\gamma}_t) = \mathcal{N}\left(\boldsymbol{\gamma}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t\right), \tag{3}$$

where the Bernoulli probabilities are parametrized using the standard normal CDF $\phi : \mathbb{R} \to (0, 1)$. The hyperparameters $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ encode the prior belief of the support for time t. Specifically, the prior mean value $\boldsymbol{\mu}_t$ controls the prior belief of the number of non-zero variables and the covariance matrix $\boldsymbol{\Sigma}_t$ determines the prior correlation of the support at time t. Thus, we can impose structure on the binary support variables $\mathbf{z}_t$ by means of imposing generic covariance functions on $\boldsymbol{\gamma}$. For example, say we choose $\Sigma_{i,j}$ to be the squared exponential covariance function, then the resulting prior distribution will promote sparsity patterns where neighbouring support variables have the same state. Under the other hand, when $\boldsymbol{\Sigma}$ is diagonal, we recover the independent spike and slab prior.

The marginal prior probability of the $x_{i,t}$ being non-zero is given by

$$\begin{aligned} p(z_{i,t} = 1) &= \int p(z_{i,t} = 1|\gamma_{i,t})p(\gamma_{i,t})\mathrm{d}\gamma_{i,t} \\ &= \int \phi(\gamma_{i,t})\mathcal{N}\left(\gamma_{i,t}|\mu_{i,t}, \Sigma_{ii,t}\right)\mathrm{d}\gamma_{i,t} \\ &= \phi\left(\frac{\mu_{i,t}}{\sqrt{1+\Sigma_{ii,t}}}\right). \end{aligned} \tag{4}$$

Thus, if the prior on $\boldsymbol{\gamma}_t$ has zero mean, then the prior belief of $p(z_{i,t})$ is unbiased, i.e. $p(z_{i,t}) = 0.5$. On the other hand, if $\mu_{i,t}$ is negative, the prior belief of $z_{i,t}$ is biased towards zero and vice versa.

III. The spatio-temporal spike and slab prior

In this section we describe the temporal extension of the structured spike and slab prior. Instead of considering $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ as fixed hyperparameters, we propose to impose a prior on $\boldsymbol{\Gamma} = \begin{bmatrix}\boldsymbol{\gamma}_1 & \boldsymbol{\gamma}_2 & \dots & \boldsymbol{\gamma}_T\end{bmatrix}$ to model problems where the support of the solution $\mathbf{X}$ changes over time. In particular, we impose a first order process Markov process on $\boldsymbol{\Gamma}$ to model the slowly changing sparsity pattern

$$p\left(\boldsymbol{\gamma}_t|\boldsymbol{\gamma}_{t-1}\right) = \mathcal{N}\left(\boldsymbol{\gamma}_t|\left(1-\alpha\right)\boldsymbol{\mu}_0 + \alpha\boldsymbol{\gamma}_{t-1}, \beta\boldsymbol{\Sigma}_0\right), \tag{5}$$

where the hyperparameters α and β control the temporal correlation and the "innovation" of the process, respectively. Furthermore, we assume that the prior distribution on $\boldsymbol{\gamma}_1$ is given by

$$p(\boldsymbol{\gamma}_1) = \mathcal{N}\left(\boldsymbol{\gamma}_1|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right). \tag{6}$$

Under these assumptions the marginal distribution of $\boldsymbol{\gamma}_2$ becomes

$$\begin{aligned} p(\boldsymbol{\gamma}_2) &= \int p\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\gamma}_1\right) p(\boldsymbol{\gamma}_1) \mathrm{d}\boldsymbol{\gamma}_1 \\ &= \mathcal{N}\left(\boldsymbol{\gamma}_2 \middle| \boldsymbol{\mu}_0, \left(\alpha^2 + \beta\right) \boldsymbol{\Sigma}_0\right). \end{aligned} \tag{7}$$

Therefore, it follows by induction that if α and β satisfy $\alpha^2 + \beta = 1$, then the marginal distribution of $\boldsymbol{\gamma}_t$ is $p(\boldsymbol{\gamma}_t) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ for all t. Furthermore, we also see that for $\alpha = 1$ and $\beta = 0$, the prior reduces to the structured spike and slab prior in the joint sparsity setting. In the other extreme, at $\alpha = 0$ and $\beta = 1$, the prior reduces to the structured spike and slab prior in the time-independent setting. Hence, the spatio-temporal spike and slab prior can be seen as a generalization of the two extreme cases.

This choice of model is also motivated by the fact that the first order structure in the temporal dimension gives rise to an inference scheme that scales linearly in the number of time steps T as we will see in the next section.

IV. Bayesian Inference using the spatio-temporal spike and slab prior

The goal of this section is to describe an inference procedure for solving the problem in eq. (1) using the proposed prior in a fully Bayesian setting. We combine the spatio-temporal spike and slab prior with a time-independent isotropic Gaussian noise model of the form

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{y}_t \middle| \mathbf{A}\mathbf{x}_t, \sigma_0^2 \mathbf{I}\right). \tag{8}$$

This gives rise to the following joint distribution

$$\begin{aligned} p(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}) = & \underbrace{\prod_{t=1}^{T} \mathcal{N}\left(\mathbf{y}_t \middle| \mathbf{A}\mathbf{x}_t, \sigma_0^2 \mathbf{I}\right)}_{f_1(\mathbf{X})} \\ & \underbrace{\prod_{t=1}^{T} \prod_{i=1}^{D} \left[(1 - z_{i,t}) \delta(x_{i,t}) + z_{i,t} \mathcal{N}\left(x_{i,t} \middle| 0, \tau_0\right)\right]}_{f_2(\mathbf{X}, \mathbf{Z})} \\ & \underbrace{\prod_{t=1}^{T} \prod_{i=1}^{D} \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\gamma_{i,t}\right)\right)}_{f_3(\mathbf{Z}, \boldsymbol{\Gamma})} \\ & \underbrace{\mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \prod_{t=2}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| (1-\alpha)\boldsymbol{\mu}_0 + \alpha \boldsymbol{\gamma}_{t-1}, \beta \boldsymbol{\Sigma}_0\right)}_{f_4(\boldsymbol{\Gamma})} \end{aligned} \tag{9}$$

The desired posterior distribution $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}|\mathbf{Y})$ is obtained from Bayes' Rule [13]. Unfortunately, this posterior distribution is intractable due to the product of mixtures and hence, we have to settle for approximate inference. Specifically, we use Expectation Propagation [14]–[16] for approximate inference by extending the proposed inference scheme in [6].

A. Approximate Inference using Expectation Propagation

Expectation propagation (EP) is an iterative deterministic method for approximating probability distributions using simpler distributions from the exponential family. As indicated in eq. (9), the exact posterior can be decomposed as follows

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}|\mathbf{Y}) = & \frac{1}{Z} \prod_{t=1}^{T} f_{1,t}\left(\mathbf{x}_t\right) \prod_{t=1}^{T} \prod_{i=1}^{D} f_{2,i,t}\left(x_{i,t}, z_{i,t}\right) \\ & \prod_{t=1}^{T} \prod_{i=1}^{D} f_{3,i,t}\left(z_{i,t}, \gamma_{i,t}\right) \prod_{t=1}^{T} f_{4,t}\left(\boldsymbol{\gamma}_t\right), \end{aligned} \tag{10}$$

where $Z = p(\mathbf{Y})$ is the normalization constant. Moreover, note that each factor in the decomposition only depends on a subset of the variables in the model, i.e. $f_{2,i,t}$ depends only on the variables $x_{i,t}$ and $z_{i,t}$ and so on and so forth. The EP framework takes advantage of this decomposition by approximating each factor in eq. (10) with a distribution from the exponential family. First we describe the functional form of the approximation and then we briefly explain how to estimate the parameters of the approximation using the EP algorithm.

Let $\tilde{f}_{1,t}$ denote the approximation of $f_{1,t}$ etc. First, we note that each of the factors in the first term, i.e. $f_{1,t}$ for all t, are already a member of the exponential family and hence does not have to be approximated. Therefore, for each t we have

$$\tilde{f}_{1,t}\left(\mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_t \middle| \tilde{\mathbf{m}}_{1,t}, \tilde{\boldsymbol{V}}_{1,t}\right), \tag{11}$$

where the parameters are determined by $\tilde{\boldsymbol{V}}_{1,t}^{-1} \tilde{\boldsymbol{m}}_{1,t} = \frac{1}{\sigma_0^2} \mathbf{A}^T \mathbf{y}_t$ and $\tilde{\boldsymbol{V}}_{1,t}^{-1} = \frac{1}{\sigma_0^2} \mathbf{A}^T \mathbf{A}$. Note that the exact term $f_{1,t}$ is a distribution on $\mathbf{y}_t$ conditioned on $\mathbf{x}_t$, whereas the approximate term $\tilde{f}_{1,t}$ is a function of $\mathbf{x}_t$ that depends on $\mathbf{y}_t$ through $\tilde{\mathbf{m}}_{1,t}$ and $\tilde{\boldsymbol{V}}_{1,t}$ etc. Next, we turn to the factors in the second term, i.e. $f_{2,i,t}$. Since each of these factors depends on $x_{i,t}$ and $z_{i,t}$, we choose $\tilde{f}_{2,i,t}$ to be

$$\tilde{f}_{2,i,t} = \mathcal{N}\left(x_{i,t} \middle| \tilde{m}_{2,i,t}, \tilde{V}_{2,i,t}\right) \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\tilde{\gamma}_{2,i,t}\right)\right), \tag{12}$$

where $\tilde{m}_{2,i,t}$, $\tilde{V}_{2,i,t}$ and $\tilde{\gamma}_{2,i,t}$ have to determined using the EP algorithm. Based on similar arguments $\tilde{f}_{3,i,t}$ and $\tilde{f}_{4,t}$ are chosen as follows

$$\tilde{f}_{3,i,t} = \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\tilde{\gamma}_{3,i,t}\right)\right) \mathcal{N}\left(\gamma_{3,i,t} \middle| \tilde{\mu}_{3,i,t}, \tilde{\Sigma}_{3,i,t}\right), \tag{13}$$

$$\tilde{f}_{4,t} = \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \tilde{\boldsymbol{\mu}}_{4,t}, \tilde{\boldsymbol{\Sigma}}_{4,t}\right). \tag{14}$$

Note that $f_{4,1}$ does not have to approximated either, it is simply $\tilde{f}_{4,1} = \mathcal{N}\left(\boldsymbol{\gamma}_1 \middle| \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$. Furthermore, note that the approximations to the factors $f_{4,t}$ for all t do not factorize w.r.t. $\gamma_{t,1}, \gamma_{t,2}, \ldots$ in order to capture potentially strong correlations in the support.

After specifying all the individual approximation terms, we derive the joint approximation of the desired posterior $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}|\mathbf{Y})$. Since the exponential family is closed under products, the approximate joint distribution has the following form

$$\begin{aligned} Q\left(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}\right) = & \prod_{t=1}^{T} \mathcal{N}\left(\mathbf{x}_t \middle| \tilde{\mathbf{m}}_t, \tilde{\boldsymbol{V}}_t\right) \prod_{t=1}^{T} \prod_{i=1}^{D} \mathrm{Ber}\left(z_{i,t} \middle| \phi\left(\tilde{\gamma}_{i,t}\right)\right) \\ & \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{\gamma}_t \middle| \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t\right). \end{aligned} \tag{15}$$

Let $\mathbf{m}_{2,t} = \left[\tilde{m}_{2,1,t}, \tilde{m}_{2,2,t}, \ldots, \tilde{m}_{2,D,t}\right]^T$ and $\boldsymbol{V}_{2,t} = \mathrm{diag}\left(\tilde{V}_{2,1,t}, \tilde{V}_{2,2,t}, \ldots, \tilde{V}_{2,D,t}\right)$, and analogously for $\tilde{\boldsymbol{\mu}}_3$, $\tilde{\boldsymbol{\Sigma}}_3$

and $\boldsymbol{\gamma}_3$, then the parameters of the joint approximation are given by

$$\tilde{\boldsymbol{V}}_t = \left(\tilde{\boldsymbol{V}}_{1,t}^{-1} + \tilde{\boldsymbol{V}}_{2,t}^{-1}\right)^{-1}, \tag{16}$$

$$\tilde{\mathbf{m}}_t = \tilde{\boldsymbol{V}}_t \left(\tilde{\boldsymbol{V}}_{1,t}^{-1}\tilde{\mathbf{m}}_{1,t} + \tilde{\boldsymbol{V}}_{2,t}^{-1}\tilde{\mathbf{m}}_{2,t}\right), \tag{17}$$

$$\tilde{\boldsymbol{\Sigma}}_t = \left(\tilde{\boldsymbol{\Sigma}}_{3,t} + \tilde{\boldsymbol{\Sigma}}_{4,t}\right)^{-1}, \tag{18}$$

$$\tilde{\boldsymbol{\mu}}_t = \tilde{\boldsymbol{\Sigma}}_t \left(\tilde{\boldsymbol{\Sigma}}_{3,t}^{-1}\tilde{\boldsymbol{\mu}}_{3,t} + \tilde{\boldsymbol{\Sigma}}_{4,t}^{-1}\tilde{\boldsymbol{\mu}}_{4,t}\right), \tag{19}$$

$$\tilde{\gamma}_{i,t} = \phi^{-1}\left[\left(\frac{\left(1-\phi(\tilde{\gamma}_{2,i,t})\right)\left(1-\phi(\tilde{\gamma}_{3,i,t})\right)}{\phi(\tilde{\gamma}_{2,i,t})\phi(\tilde{\gamma}_{3,i,t})} + 1\right)^{-1}\right]. \tag{20}$$

The posterior covariance matrices $\tilde{\boldsymbol{V}}_t$ and $\tilde{\boldsymbol{\Sigma}}_t$ are (potentially) fully dense matrices, which makes the approximation able to cope with non-orthogonal forward matrices $\mathbf{A}$.

B. The Expectation Propagation Algorithm

In this section we describe how to compute the parameters of the individual approximations using the EP algorithm. The EP algorithm works by updating each of the individual approximation terms one by one until convergence. Consider the update of the term $\tilde{f}_{a,i,t}$ for a given a, i and t. The update is obtained by performing the following three steps of the EP algorithm. The first step is to remove the contribution of $\tilde{f}_{a,i,t}$ from the joint approximation in eq. (15) by forming the so-called cavity distribution

$$Q^{\setminus a,i,t} \propto \frac{Q}{\tilde{f}_{a,i,t}}. \tag{21}$$

In the next step we minimize the Kullbach-Leibler [13] divergence between $f_{a,i,t}Q^{\setminus a,i,t}$ and $Q^{a,t,\text{new}}$ w.r.t. $Q^{a,t,\text{new}}$. That is, we minimize $\text{KL}\left(\frac{1}{Z_{a,i,t}}f_{a,i,t}Q^{\setminus a,i,t}||Q^{a,t,\text{new}}\right)$, where $Z_{a,i,t}$ is the normalization constant of $f_{a,i,t}Q^{\setminus a,i,t}$. For distributions within the exponential family, minimizing this form of KL divergence amounts to matching moments between $f_{a,i,t}Q^{\setminus a,i,t}$ and $Q^{a,t,\text{new}}$ [14]. Finally, the third and last step is to compute the new update of $\tilde{f}_{a,i,t}$ as follows

$$\tilde{f}_{a,i,t} \propto \frac{Q^{a,t,\text{new}}}{Q^{\setminus a,i,t}}. \tag{22}$$

After the individual approximation terms $\tilde{f}_{a,i,t}$ for all i and t for a given a have been updated, the relavant parts of the joint approximation are updated using eq. (16)-(20). To minimize the computational load, we use parallel updates of $\tilde{f}_{2,i,t}$ [9] followed by parallel updates of $\tilde{f}_{3,i,t}$ rather than the conventional sequential update scheme. Furthermore, due to the fact that $\tilde{f}_2$ and $\tilde{f}_3$ factorizes w.r.t. both i and t, we only need the marginals of the cavity distributions $Q^{\setminus a,i,t}$, which simplifies the computations. Computing the cavity distributions and matching the moments are straightforward. However, when matching the moments, we are required to evaluate of the zero'th, first and second order moment of the distributions of the form $\phi(\gamma_i)\mathcal{N}\left(\gamma_i\middle|\mu_i,\Sigma_{ii}\right)$. Derivation of analytical expressions for these moments can be found in the appendix to chapter 3 in [17].

The proposed EP algorithm is summarized in figure 1. The computational complexity of the algorithm is dominated by the matrix inversions in eq. (16) and (19). However, when $N < D$, the covariance matrices $\tilde{\boldsymbol{V}}_{1,t}$ have low rank and hence, eq. (16) can be carried out in $\mathcal{O}\left(ND^2\right)$ using the Matrix Inversion Lemma [18]. Therefore, the resulting inference scheme scales as $\mathcal{O}\left(TD^3\right)$, i.e. it scales linearly in the number of measurement vectors T.

- Initialize approximation terms $\tilde{f}_a$ for $a = 1, 2, 3, 4$ and Q
- Repeat until stopping criteria
  - For each $\tilde{f}_{2,i,t}$:
    - Compute cavity distribution: $Q^{\setminus 2,i,t} \propto \frac{Q}{\tilde{f}_{2,i,t}}$
    - Minimize: $\text{KL}\left(f_{2,i,t}Q^{\setminus 2,i,t}\middle||Q^{2,t,\text{new}}\right)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{2,i,t} \propto \frac{Q^{2,t,\text{new}}}{Q^{\setminus 2,i,t}}$ to update parameters $\tilde{m}_{2,i,t}, \tilde{v}_{2,i,t}$ and $\tilde{\gamma}_{2,i,t}$.
  - Update joint approximation parameters: $\tilde{\mathbf{m}}, \tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{\gamma}}$
  - For each $\tilde{f}_{3,i,t}$:
    - Compute cavity distribution: $Q^{\setminus 3,i,t} \propto \frac{Q}{\tilde{f}_{3,i,t}}$
    - Minimize: $\text{KL}\left(f_{3,i,t}Q^{\setminus 3,i,t}\middle||Q^{3,t,\text{new}}\right)$ w.r.t. $Q^{3,t,\text{new}}$
    - Compute: $\tilde{f}_{3,i,t} \propto \frac{Q^{3,t,\text{new}}}{Q^{\setminus 3,i,t}}$ to update parameters $\tilde{\mu}_{3,i,t}, \tilde{\sigma}_{3,i,t}$ and $\tilde{\gamma}_{3,i,t}$
  - Update joint approximation parameters: $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\gamma}}$
  - For each $\tilde{f}_{4,t}$
    - Compute cavity distribution: $Q^{\setminus 4,t} \propto \frac{Q}{\tilde{f}_{4,t}}$
    - Minimize: $\text{KL}\left(f_{4,t}Q^{\setminus 4,i}\middle||Q^{4,t,\text{new}}\right)$ w.r.t. Q^{new}
    - Compute: $\tilde{f}_{4,t} \propto \frac{Q^{4,t,\text{new}}}{Q^{\setminus 4,t}}$ to update parameters $\tilde{m}_{4,t}, \tilde{v}_{4,t}$ and $\tilde{\gamma}_{4,t}$.
  - Update joint approximation parameters: $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}$

Fig. 1. Proposed algorithm for approximating the joint posterior distribution over $\mathbf{X}$, $\mathbf{Z}$ and $\boldsymbol{\Gamma}$ conditioned on $\mathbf{Y}$.

C. Tuning of hyperparameters

The algorithm requires tuning of multiple hyperparameters for optimal performance. The Expectation Propagation framework provides a neat alternative to typical cross-validation schemes. Besides the approximation to the posterior distribution $P(\mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}\big|\mathbf{Y})$, EP also provides an approximation to the marginal likelihood $P(\mathbf{Y})$, which is very useful for model selection and tuning of hyperparameters [13]. The exact marginal likelihood is obtained by marginalizing out $\mathbf{X}, \mathbf{Z}$ and $\boldsymbol{\Gamma}$ from the joint distribution in eq. (9). The EP approximation to the marginal likelihood is obtained by substituting all the (scaled) individual approximation terms into the resulting formula. Finally, it is also possible to get closed form expression for the gradients of the marginal likelihood approximation w.r.t. to the hyperparameters [16], [17], which allows efficient tuning of the hyperparameters. However, a detailed treatment of the marginal likelihood approximation and its gradient w.r.t. hyperparameters are out of scope for this extended abstract.

V. Numerical Experiments

In order to investigate the properties of the proposed algorithm, we have designed and conducted two numerical experiments. The first experiment addresses the reconstruction performance of the algorithm, whereas the second experiment investigate the algorithm's robustness towards coherent forward models.

A. Experiment 1

To evaluate the proposed method, we have compared the method to several related solvers: BG-AMP[1] [19], DCS-AMP[2] [20], Spatial

[1]We used the implementation in GAMP-toolbox by Sundeep Rangan et al: http://gampmatlab.wikia.com/wiki/

[2]We used the implementation in the DCS-AMP-toolbox by Justin Ziniel: http://www2.ece.ohio-state.edu/∼zinielj/dcs/

EP (implements the structured spike and slab prior) [6] and Spatial MMV EP. The BG-AMP method combines the conventional spike and slab prior with approximate message passing-based [21] inference. We include this method to have a baseline result without any structural assumptions on the sparsity pattern. The DCS-AMP can be seen as an extension of BG-AMP, which assumes that the sparsity pattern evolves slowly in time according to a binary Markov chain. The Spatial EP method assumes spatial correlation in the sparsity pattern, but no temporal correlation. Finally, the Spatial MMV method is similar to Spatial EP but with static sparsity across time, i.e. it assume joint sparsity across time.

To set up the first test we first sampled one realization of $\mathbf{Z}$ using eq. (2)-(5) with $D = 100$, $T = 100$, $\alpha = 0.99$ and $\beta = 1 - \alpha^2$, see figure 4(a). The average number of non-zero weights per column is fixed to 20. We note that the resulting sample exhibits the spatio-temporal structure as desired. Afterwards, we sample the nonzero coefficients in $\mathbf{X}$ from a standard normal distribution and from these we generate compressive measurements using eq. (1), where $A_{ij} \sim \mathcal{N}(0, 1/N)$, the SNR $= 10$dB and the undersampling ratio N/D is varied from from 0.05 to 0.95. To quantify the performance of the methods we use Normalized Mean Square Error (NMSE) between the true $\mathbf{X}$ and the estimated $\hat{\mathbf{X}}$ given by

$$NMSE = \frac{\sum_{i,t}\left(X_{i,t} - \hat{X}_{i,t}\right)^2}{\sum_{i,t} X_{i,t}^2}. \tag{23}$$

Furthermore, we evaluate each method's ability to recover the true support $\mathbf{Z}$ using the F-measure [22] based on a MAP estimate of the support $\hat{\mathbf{Z}}$,

$$F = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{24}$$

The results are averaged over 100 realizations of the noise $\mathbf{E}$ and non-zero coefficients in $\mathbf{X}$ and are shown in figures 2-3. It is seen that the proposed spatio-temporal method outperforms the other methods both in terms of NMSE and F-measure, but in general it is seen that richer prior assumptions on the support improves the results significantly. We also note that for very undersampled problems, the Spatial MMV EP method with static sparsity actually performs best. But as the undersampling ratio increases, all the other methods, including BG-AMP, outperforms it due to the very high bias of the model.

Figures 4(b)-4(f) shows the reconstructed support sets for the undersampling ratio $N/D = 0.4$. It is seen that DCS-AMP and Spatial EP, which models temporal and spatial structure, respectively, clearly outperforms BG-AMP. Furthermore, it is also seen that joint sparsity assumption (fig. 4(e)) is too restrictive for these kinds of signals. Again, we note that the spatio-temporal model gives superior results in terms of both F-measure and NMSE.

B. Experiment 2

The forward model $\mathbf{A}$ in the EEG source localization problem contains highly correlated columns, i.e. $\mathbf{A}$ is coherent. Therefore, it is of interest to investigate the proposed algorithm's robustness to coherent forward models. The set-up in this experiment is basically the same as for the first experiment, except that undersampling ratio is now fixed to $N/D = 0.4$ and the elements in the forward model A_{ij} are no longer Gaussian i.i.d. Instead we sample the rows of $\mathbf{A}$ from a correlated multivariate normal distribution, such that the

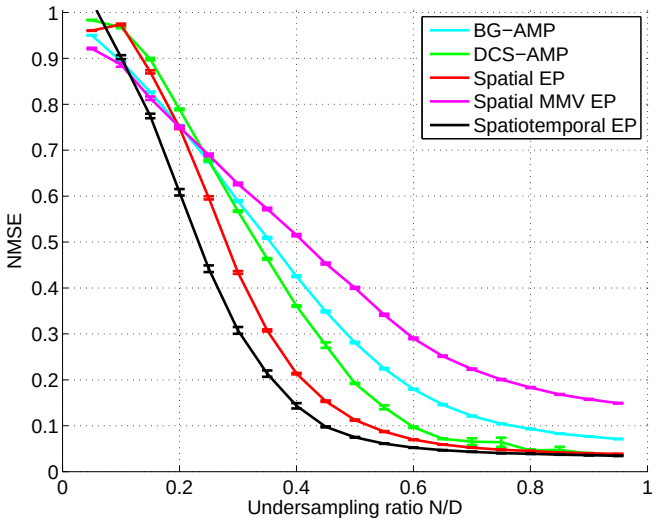

Fig. 2. Normalized mean square error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a), where $D = 100, T = 100$ and SNR $= 10dB$. The entries in $\mathbf{A}$ are Gaussian i.i.d, i.e. $A_{i,j} \sim \mathcal{N}(0, 1/N)$. The results are averaged over 100 realizations.

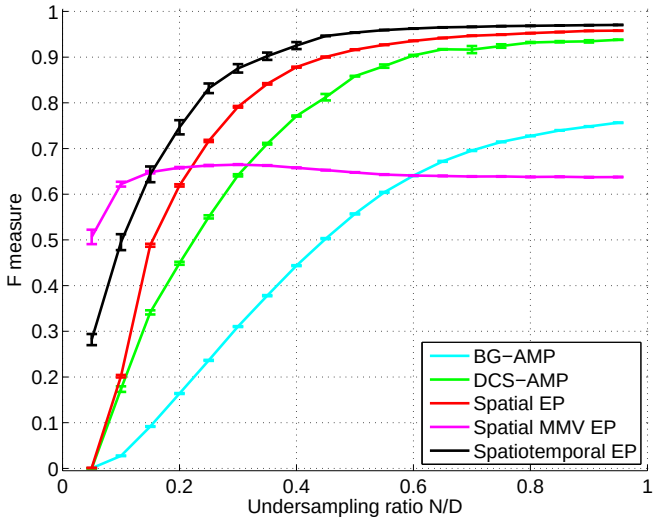

Fig. 3. F-measure error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a), where $D = 100, T = 100$ and SNR $= 10dB$. The entries in $\mathbf{A}$ are Gaussian i.i.d, i.e. $A_{i,j} \sim \mathcal{N}(0, 1/N)$. The results are averaged over 100 realizations.

columns of $\mathbf{A}$ will be correlated. In particular, the correlation of the i'th and j'th column of $\mathbf{A}$ is given by $r^{|i-j|}$. We compute the NMSE and F-measure as a function of the correlation r. Note that the BG-AMP and DCS-AMP methods are designed for Gaussian i.i.d forward. These two methods are therefore not expected to perform well in this experiment, but we include them for completeness. The results are averaged over 50 realizations and are shown in figures 5 and 6. The EP-based methods show some robustness to correlation in the columns of $\mathbf{A}$, but the performance does degrade gradually when we increase the correlation. In particular, when changing the correlation r from 0.05 to 0.95, the F-measure for the spatio-temporal method only drops from approximate 0.92 to 0.89, but the NMSE increases from approximately 0.15 to 0.45.

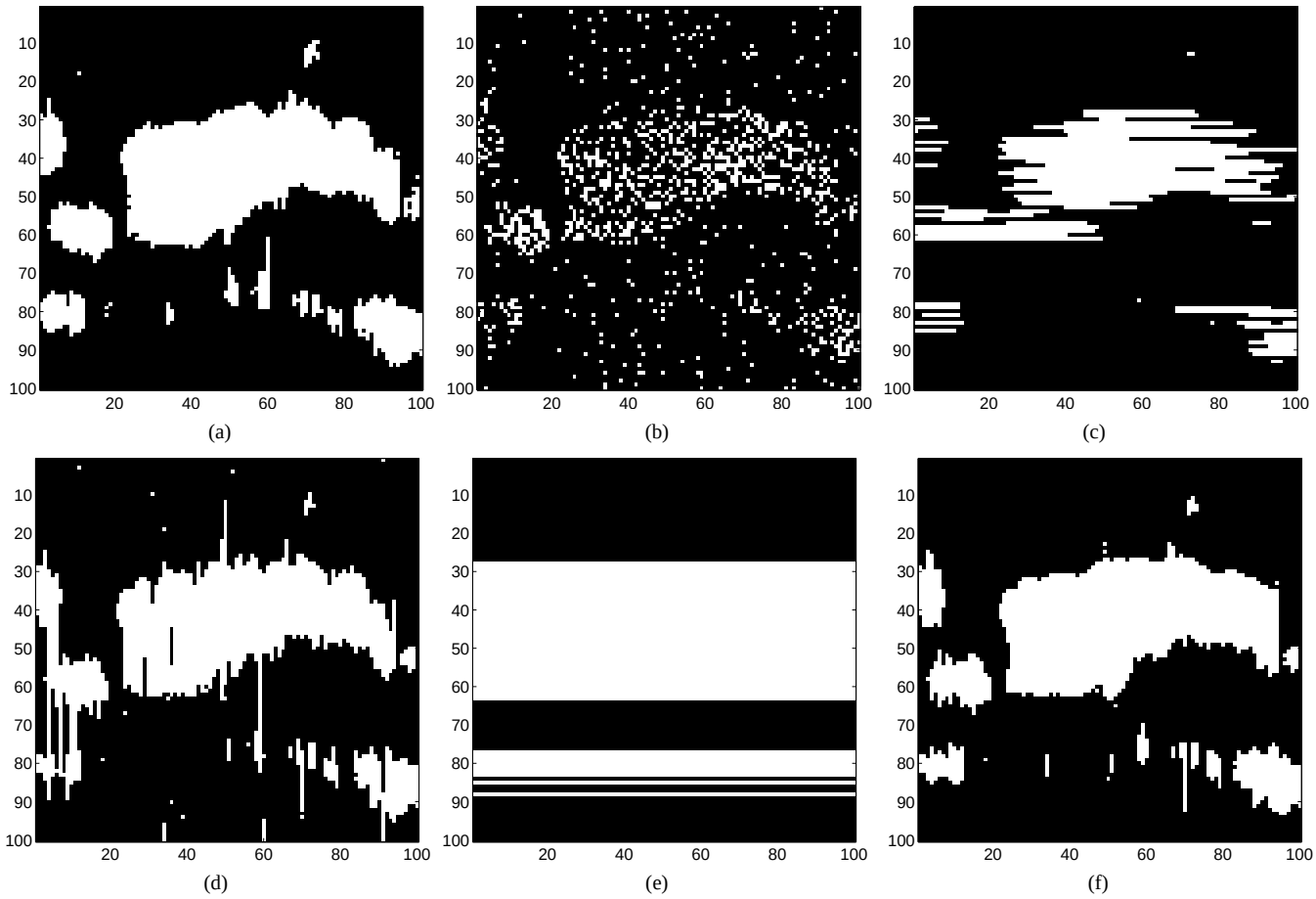

Fig. 4. True and reconstructed support for the 5 considered methods. The undersampling ratio is $N/D = 0.4$ and $D = 100, T = 100$ and $SNR = 10$dB. a) True support, b) BG-AMP (NMSE = 0.805, F = 0.450), c) DCS-AMP (NMSE = 0.777, F = 0.763), d) Spatial EP (NMSE = 0.658, F = 0.902), e) Spatial MMV EP (NMSE = 0.833, F = 0.663), f) Spatio-temporal EP (NMSE = 0.618, F = 0.935).

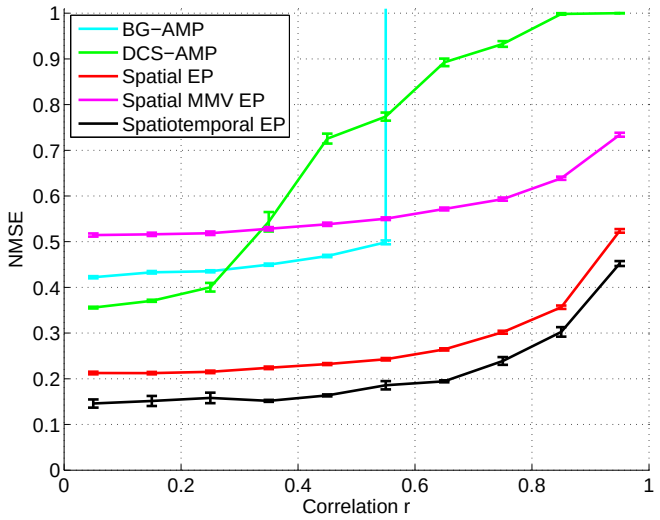

Fig. 5. NMSE error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a). The correlation of the i'th and j'th column of $\mathbf{A}$ is given by $r^{|i-j|}$. The results are averaged over 50 realizations.

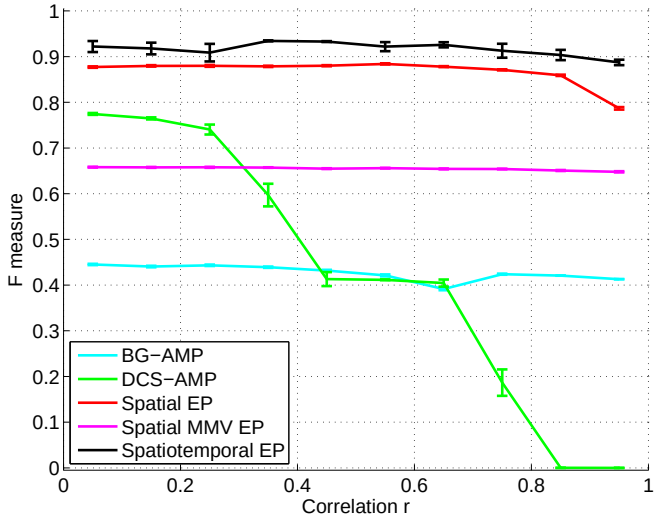

Fig. 6. F-measure error as a function of undersampling ratio. The data are generated from $\mathbf{Y} = \mathbf{AX} + \mathbf{E}$ with the sparsity pattern shown in figure 4(a). The correlation of the i'th and j'th column of $\mathbf{A}$ is given by $r^{|i-j|}$. The results are averaged over 50 realizations.

VI. Conclusion & outlook

We extended the structured spike and slab prior and the associated Expectation Propagation inference scheme to cope with smooth temporal evolution of the sparsity pattern. Based on numerical experiments with synthetic data we demonstrated the benefits of the extended model. In particular, we showed that the method outperformed the reference methods. Future work includes developing an automated approach learning the hyperparameters of the prior and applying the proposed method to a real EEG source localization problem.

Acknowledgment

The authors would like to thank Sundeep Rangan et al. and Justin Ziniel for making their toolboxes available online.

References

[1] S. F. Cotter, B. D. Rao, K. Engan, K. Kreutz-delgado, and S. Member, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, pp. 2477–2488, 2005.

[2] S. Baillet, J. C. Mosher, and R. M. Leahy, "Electromagnetic brain mapping," *IEEE Signal Processing Magazine*, vol. 18, no. 6, pp. 14–30, 2001.

[3] J. M. Antelis and J. Minguez, "EEG source localization based on dynamic bayesian estimation techniques," 2012.

[4] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, "Sparse signal recovery using markov random fields," *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, pp. 257–264, 2009.

[5] E. van den Berg and M. P. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Transactions On Information Theory*, vol. 56, no. 5, pp. 2516–2527, 2010.

[6] M. R. Andersen, O. Winther, and L. K. Hansen, "Bayesian inference for structured spike and slab priors," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1745–1753.

[7] L. Jacob, J.-P. Vert, G. Obozinski, and G. Obozinski, "Group lasso with overlap and graph lasso," *Proceedings of the 26th International Conference on Machine Learning, ICML 2009*, pp. 433–440, 2009.

[8] T. J. Mitchell and J. Beauchamp, "Bayesian variable selection in linear-regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[9] D. Hernandez-Lobato, J. Hernandez-Lobato, and P. Dupont, "Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation," *Journal Of Machine Learning Research*, vol. 14, pp. 1891–1945, 2013.

[10] L. Yu, H. Sun, J. P. Barbot, and G. Zheng, "Bayesian compressive sensing for clustered sparse signals," *Icassp, Ieee International Conference on Acoustics, Speech and Signal Processing - Proceedings, Icassp Ieee Int Conf Acoust Speech Signal Process Proc*, pp. 3948–3951, 2011.

[11] D. Donoho, "Compressed sensing," *IEEE TRANSACTIONS ON INFORMATION THEORY*, vol. 52, no. 4, pp. 1289–1306, 2006.

[12] J. Ziniel, L. C. Potter, and P. Schniter, "Tracking and smoothing of time-varying sparse signals via approximate belief propagation," *Conference Record - Asilomar Conference on Signals, Systems and Computers, Conf. Rec. Asilomar Conf. Signals Syst. Comput*, pp. 808–812, 2010.

[13] C. M. Bishop, "Pattern recognition and machine learning," 2006.

[14] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, ser. UAI '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[15] M. Opper and O. Winther, "Gaussian processes for classification: Mean-field algorithms," *Neural Computation*, vol. 12, no. 11, pp. 2655–2684, 2000.

[16] M. Seeger, "Expectation propagation for exponential families," 2009.

[17] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[18] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2012.

[19] J. P. Vila and P. Schniter, "Expectation-maximization gaussian-mixture approximate message passing," *IEEE Transactions On Signal Processing*, vol. 61, no. 19, pp. 4658–4672, 2013.

[20] J. Ziniel and P. Schniter, "Dynamic compressive sensing of time-varying signals via approximate message passing," *IEEE transactions on signal processing*, vol. 61, no. 21, pp. 5270–5284, 2013.

[21] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," *Ieee International Symposium on Information Theory - Proceedings, Ieee Int Symp Inf Theor Proc*, pp. 2168–2172, 2011.

[22] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.